The SOM reef - A new metaphoric visualization approach for self organizing maps

Tim W. Nattkemper
Applied Neuroinformatics Group, Faculty of Technology, Bielefeld University
Bielefeld, Germany
tnattkem@techfak.uni-bielefeld.de

Abstract - *In this paper a new tool for visualizing self organizing maps (SOM) for interdisciplinary explorative data analysis is proposed, the SOM reef. By combination of the U-matrix visualization approach with the new introduced fish glyph, the SOM can be displayed as a underwater scenario. Data sets can be visualized in the SOM as swarms of fishes or prototypes. The SOM reef provides (i) a powerful metaphor for comunicating SOM results and (ii) an entertaining interface for exhaustive data exploration.*

Key words - Information Visualization, Explorative Data Analysis, Data Glyphs

1 Introduction

The aim of data mining and knowledge discovery is to search large sets of N-dimensional data for hidden regularities, clusters, or other relations. Since humans have great visual skills, visualization of the data for the purpose of an interactive analysis by a human observer is a promissing approach and has been stimulating scientific and engineering research.
But although Ben Shneiderman's famous mantra for interactive visual data exploration (*"Overfiew first, zoom and filter, then details on demand"*) [1] is convincing from the point of view of visualizatiuon engineers as well as for psychophysicists, who analyze human-computer interaction, there is still no straightforward sheme for designing information visualization systems for N-dimensional data. And interestingly, scientific textbooks about information visualization have been published majorly in the last six years [2, 3, 4, 5, 6].
One straightforward visualization approach to analyze N variables in m observations is to map the variables to the attributes (in general shape, size, colour and location) of graphically displayed entities, so called *glyphs*. However, the comprehensive glyph display for the entire set of m samples has just been identified as an interesting scientific engineering problem [7]:" *The placement or layout of glyphs on a display can communicate significant information regarding the data values themselves as well as relationships between data points, ...*". In this regard, self-organizing maps (SOM) (and other techniques for dimensional reduction and projection) offer interesting solutions to this problem and has been established as a tool for visual data exploration. To visualize the trained SOM, several approaches have been proposed: The feature density of the trained SOM prototype vectors is displayed based on smoothed histograms [8] or the U-matrix [9] apply clustering to the set of prototype vectors [10, 11]. For the special case of very large SOMs, fish eye view or fractal view have been

proposed [12]. In addition, the SOM visualization can be augmented by text labels, as for instance the WEBSOM [13] or a single feature analysis with a component plane view [14]. Also automatic feature selection has been proposed to render icons for displaying the SOM prototype vectors on a grid [15].
SOM are used in many interdisciplinary projects like biopattern analysis, customer data mining in economics or providing convinient interfaces in text mining applicati ons. So the communication of the result SOM is a substantial step in interdisciplinary discussion. But we observed that explaining the meaning of a SOM visualization is often time consuming or even frustrating. One reason for this is that explaining the SOM contains a lot of standard vocabular from the fields of algebra, pattern recognition or artifical neural networks, sometimes unknown to the research collaborators from field of economics or biomedicine. In fact we often observed that the partners had interpretations of terms like *pattern*, *vector* or *similarity* which were different from the concepts in the fields of pattern recognition or artificial neural networks.
One of the most powerful tools for explaining structures or relations to people with a different background knowledge is the *metaphor*, i. e. to compare two seemingly unrelated subjects. Its explanatory power lies in the opportunity to describe one subject (the SOM) by the comparison with a familier real world subject, well known to both parties (like a natural scene). The basic idea behind this work was to design a metaphoric SOM visualization tool, where the data structure is interpreted as a cartoon for a natural scene, in this case a underwater scenario with a reef full of fishes. An approach for computing a metaphoric description of a projection result would be of valuable help to the computer scientist to discuss his results with his collaborative partners from biology, chemistry etc.
Another motivation for this work was the fact, that interpreting a trained SOM is difficult regarding the identification of features that have influence in the vector quantization. This problem gets serious for SOMs of large numbers of nodes and/or a large data dimension. In addition one has to consider, that in most data mining projects the data may have *several* structural features to be discovered. Especially the analysis of heterogenous clusters and outliers can be time consuming. So information analysts may need to spend some time with the data and its visualizations which can be quite tiresome and boring. In this case one would benefit from displays, that catch the attention of the user again and again. This favourable display quality could be called *entertaining*.

2 The SOM reef tool

The SOM reef visualization tool renders an underwater cartoon as a metaphor for a SOM result on two steps. In a first step, after training the SOM with a training set $\Gamma_t = \{\mathbf{x}_\alpha\}$ (with $\mathbf{x}_\alpha \in [0;1]^N$ and $\alpha = 1, \ldots, m$) the trained SOM prototype vectors are used to render the seabed profile using the U-matrix approach [9]. In a second step, the explored data $\Gamma_e = \{\mathbf{x}_\beta\}$ with $\mathbf{x}_\beta \in [0;1]^N$ is mapped on top of the seabed using the new developed fish glyph. In this work we consider the typical exploration case $\Gamma_t = \Gamma_e$.

2.1 The U-matrix

The U-matrix has been proposed in [9] and is probably the most applied visualization framework for SOM, especially for SOM with a large number of neurons. The U-matrix displays

the densities in the feature space across the SOM grid. To this end, pairwise distances between SOM node prototype vectors are computed and arranged in a low-dimensional array at positions corresponding to the grid node positions. These intensities are displayed by a height profile or by a colored plane (or by both). Thus, the U-matrix itself can already give a metaphoric description of the data density by an image of mountains. In this work we visualize the U-matrix in a hybrid manner as a colorized height profile. We use a color scale which has been adjusted manually to simulate the color changes of the seabed depending on the depth, i. e. a scale from *cyan* to *blue* to *black*. In most applications the U-matrix is displayed as a height profile, with the height being proportional to the distance between prototype vectors. So in the display clusters of very different features are separated by a ridge of mountains. Since we consider an underwater scenario we visualize the U-matrix the other way round, i. e. we draw the *depths* of the seabed proportional to the feature distances. In addition to color and depth the system can also render texture in the U-matrix to support the visual interpretation of the landscape profile. To this end we generated a set of twelve static texture patches $T_0, \ldots T_{11} \in 128 \times 128$. The texture patches T_i are generated be randomly drawing two-dimensional grey valued Gaussians. The number and the amplitude of the Gaussians are constantly increased from T_0 to T_{11} (see Fig. 1 B)).

2.2 The Fish Glyph

To render the protoptype vectors and vectors from a training or test set $\Gamma_{(t,e)}$ which are projected on the SOM, we propose a new N-dimensional data glyph. In general, glyphs can be categorized as *abstract* or *metaphoric*. Abstract glyphs are basic geometric models without direct symbolic or semantic interpretration like profiles [16], stars [17], boxes [18]. To display more variables or also data relations, abstract glyphs can get quite complex like the customized glyphs [19, 20], shapes [21] or infochrystals [22]. Such glyphs are powerful tools for a compact display of variable relations. However, the user must spend considerable time for training to be able to use these tools effectively.
The first idea for a metaphoric display led to the introduction of the well known Chernoff faces [23] The idea of rendering data faces may get new stimuli from advances in computer graphics and animation [24] since a large range of algorithms exist to render faces in different emotional states. However, the succesful application of Chernoff faces seems to be restricted to data with a one-dimensional substructure, like social and economic parameters as in [25, 26, 27]. Similar approaches use stick figures [28], a parametrized tree [29] or wheels [30].
To visualize the SOM in a metaphoric manner, we need to synchronize the designs of the U-matrix landscape and the data glyphs. To this end we developed a fish shaped glyph that can be rendered on top of the U-matrix seabed. We voted for an underwater scenario for several reasons. First, underwater scenarios gained some popularity and are increasingly distributed across all kinds of media, from motion pictures to documentary films on TV. Second, fish live as loners as well as in swarms, thus clusters of fish as well as single fish appear as natural. Third, fishes have lots of features that can be parametrized straightforward and easily. In biology, fish species are generally grouped in clusters according to their physiological features (like *blue whiptail, paradise whiptail, double whiptail* and so on) and each cluster can be represented by a prototype (i. e. *whiptail* [31]) similar to the idea of vector quantization. In this first software prototype, the fish is rendered based on a grid model which has 17 graphical attributes. They consist of 14 geometric attributs (6 angles and 8 arc length) and 4

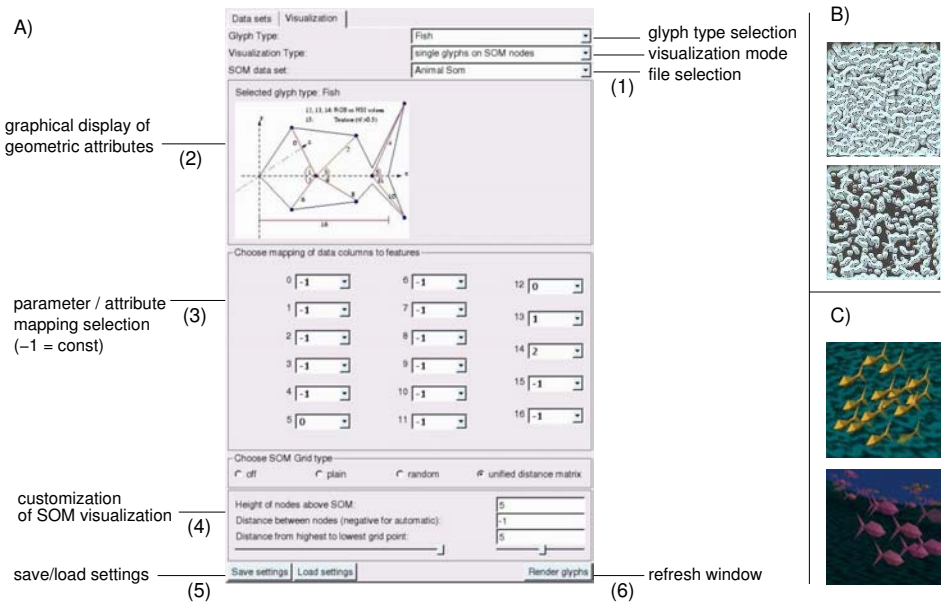

Figure 1: **A):** The SOM reef interface: In (1) the user selects a glyph type (in this first version only the fish glyph is realized), a visualization type (glyphs on SOM is default, the tool can also plot fishes in a scatterplot fashion) and a data set. In (2) the geometric and appearance attributes are displayed, so the user can easily map variables to attributes in (3). In this example, the first three variables are chosen to set the color of the fish (attribute 12, 13, 14) and the first variable is also mapped to one angle of the caudal fin (at the end of the fish). In (4) the distance between (i) the U-matrix landscape and the glyphs and (ii) the highest and lowest point in the landscape can be set. The user can also choose, how the fishes can be rendered around their best match node. The buttons in (5) activate loading and saving of good parametrizations and (6) acivates a new rendering of the SOM reef based on new parameters. **B):** Four examples from the texture patch set T_0, T_{11}. **C):** Two examples of fish glyphs.

appearence attributes (RGB color plus texture) as displayed in Fig.1. As already summerized in [32, 7], humans' abilities for percieving graphical attributes of glyphs vary considerably. Thus, the software is designed to allow a convinient customization of mapping variables to graphical parameters.

3 Applications and Results

The SOM reef is computed and displayed for two data sets.

3.1 The Iris data set:

This data set is one of the standard benchmark data sets in machine learning and pattern classification. It contains a data set with 150 random samples of flowers from the *iris species setosa, versicolor,* and *virginica*. From each species there are 50 observations for the features *sepal length, sepal width, petal length*, and *petal width* in cm. The data set is used to train a 10×10 SOM. In Fig. 2 the SOM reef of the iris data set is shown. In this case the seabed is visualized without texture. The four variables are mapped to geometric attributes, the prototype vectors are plotted with 50% transparency, the original data Γ_e is displayed as solid.

3.2 The COIL data set:

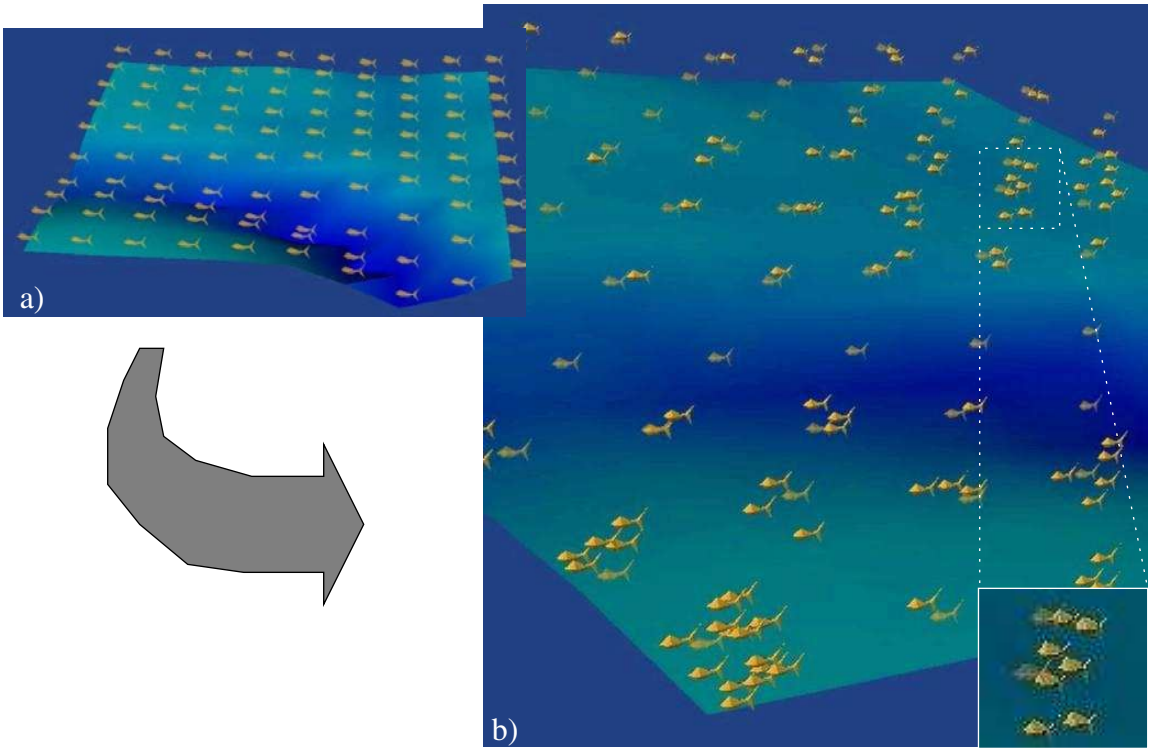

Figure 2: The 10×10 SOM of the iris data set is displayed. The U-matrix is visualized using the manually derived color scale. The four variables are mapped to geometrical attributes of the glyphs. The prototype vectors are displayed transparent. The data set is projected onto the SOM as solid fish glyphs. The SOM is partitioned into two regions divided by a canyon in the lower left corner. The fish glyphs in both regions differ in shape.

The columbia image library (COIL) provides images of 20 different objects viewed form different directions [33]. On the entire data set a principal component analysis (PCA) is performed. The eigenvectors of the ten largest eigenvalues account for most of the signal intensity variance and are used to project each image to a ten dimensional vector. A 50×50 SOM is trained with this set and the result SOM is displayed as a SOM reef in Fig. 3.

4 Discussion

A new approach for SOM visualization has been proposed. In contrast to other works, the approach aims at a metaphoric explanation of the SOM to non-expert observers. The metaphoric display consists of visualizing the SOM U-matrix as an underwater seabed using color and texture plus rendering single feature vectors as fish shaped glyphs. The glyph interface allows easy and convenient mapping of variables to glyph parameters. The examples show, that shape and color of the fishes can represent feature variables and the appealing look of the SOM reef. We believe that the SOM reef will improve SOM based data analysis by (a) making the SOM inspection more entertaining and (b) providing easy-to-interpret metaphoric SOM display for non-expert users. Future work will aim at advancing the reef scenario and developing alternative scenarios like a field of flowers.

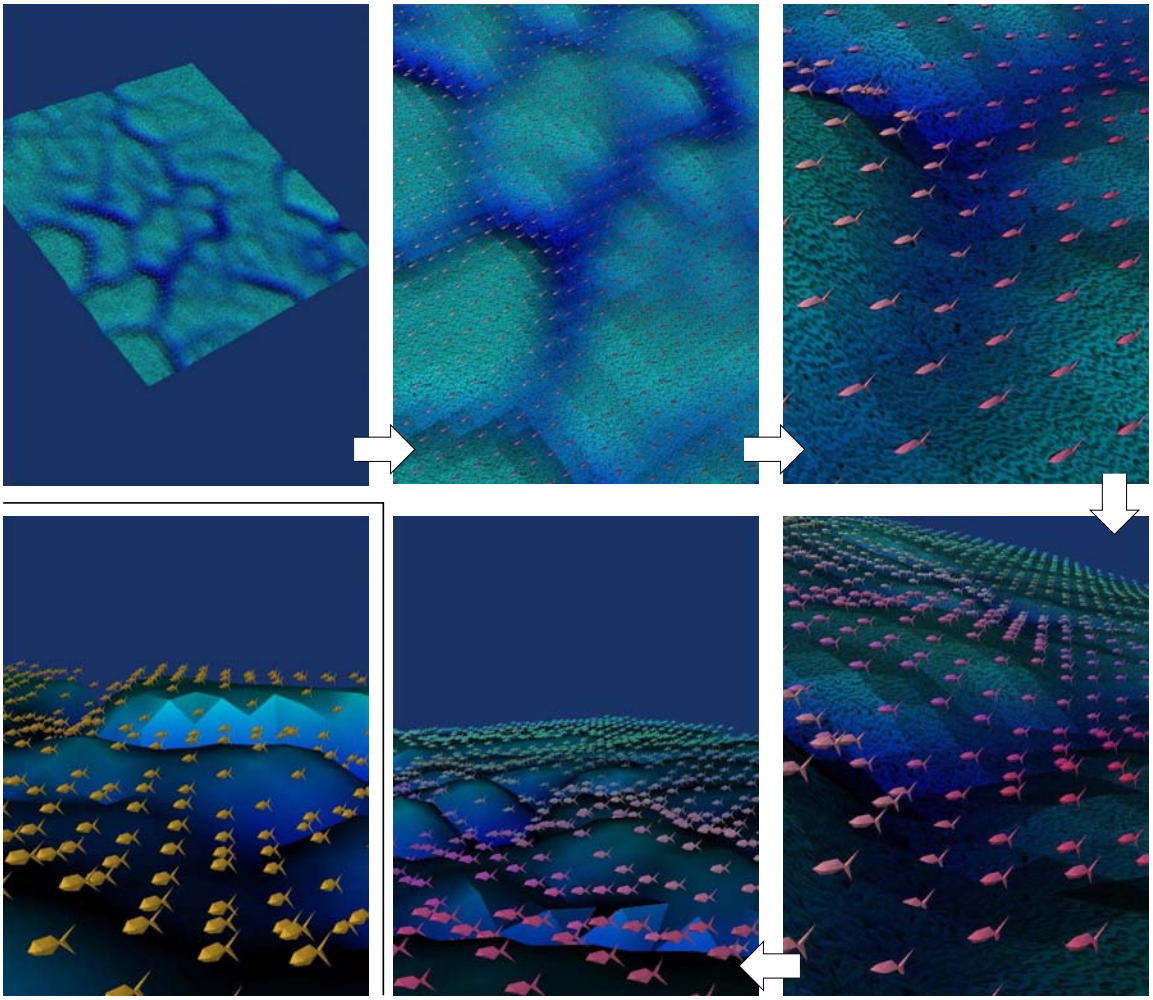

Figure 3: The 50×50 SOM reef of the coil data set. The ten variables of the data set are mapped to geometric attributes of the fish glyphs, the first three variables are mapped to the RGB values of the fishes. The U-matrix is visualized by color scale and texture mapping. In the separated image on the lower left the variables are mapped solely to shape parameters to show how the fish shapes change across the SOM.

Acknowledgement: Special thanks go to Harmen Grosse Deters and Wiebke Timm for computational support and helpful comments.

References

[1] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualization. In *Proc. of the IEEE Symp. on Vis. Lang.*, pages 336–43, 1996.

[2] S.K. Card, J.D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization.* Morgan Kaufmann Publishers, 1999.

[3] Robert Spence. *Information Visualization.* Addison Wesley Longman, 2000.

[4] U. Fayyad, G.G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann Publishers, 2001.

[5] C. Ware. *Information Visualization*. Morgan Kaufmann Publishers, 2004.

[6] C. Chen. *Information Visualisation & Virtual Environments*. Springer, 2004.

[7] M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1:194–210, 2002.

[8] J. Vesanto. Som-based visualization methods. *Intelligent Data Analysis*, 3:111–126, 1999.

[9] A. Ultsch. Self organizing neural networks for visualization and classification. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 307–13. Springer, 1993.

[10] J. Vesanto and E. Alhoneimi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11:586–600, 2000.

[11] S. Wu and T. W. S. Chow. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, 37:175–188, 2004.

[12] C. C. Yang, H. Chen, and K. K. Hong. Visualization tools for self-organizing maps. In *Proc. of the 4th ACM conf. on Digital libraries*, pages 258–9, 1999.

[13] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom - self-organizing maps of document collections. In *Proc. of WSOM*, 1997.

[14] S. Kaski, J. Nikkilä, and T. Kohonen. Methods for interpreting a self-organized map in data analysis. In *Proc. of ESANN*, 1998.

[15] A. Rauber and D. Merkl. Automatic labeling of self-organizing maps for information retrieval. *Journal of Systems Research and Information Systems (JSRIS)*, 10(10):23–45, December 2001.

[16] S. du Toit, A. Steyn, and R. Stumpf. *Graphical exploratory data analysis*. Springer, 1986.

[17] J. Siegel, E. Farrell, R. Goldwyn, and H. Friedman. The surgical implication of physiologic patterns in myocardial infarction shock. *Surgery*, 72:126–41, 1972.

[18] J. Hartigan. Printergraphics for clustering. *ournal. of Statistical Computing and Simulation*, 4:187–213, 1975.

[19] M.W. Ribarsky, E. Ayers, J. Eble, and S. Mukherjea. Glyphmaker: Creating customized visualizations of complex data. *IEEE Computer*, 27(7):57–64, July 1994.

[20] Martin Kraus and Thomas Ertl. Interactive data exploration with customized glyphs. In V. Skala, editor, *WSCG 2001 Conference Proceedings*, 2001.

[21] Christopher D. Shaw, James A. Hall, Christine Blahut, David S. Ebert, and D. Aaron Roberts. Using shape to visualize multivariate data. In *Workshop on New Paradigms in Information Visualization and Manipulation*, pages 17–20, 1999.

[22] A. Spoerri. Infocrystal: a visual tool for information retrieval & management. In *Proceedings of the second international conference on Information and knowledge management*, Washington, D.C., United States, 1993. ACM Press.

[23] H. Chernoff. The use of faces to represent points in n-dimensional pace graphically. Technical Report RN NR-042-993, Dept. of Stat., Stanford Univ., 1971.

[24] J.-y. Noh and U. Neumann. A survey of facial modeling and animation techniques. Technical Report 99-705, USC Technical Report, 1998.

[25] D. Dorling. Cartograms for visualizing human geography. In H. M. Hearnshaw and D. J. Unwin, editors, *Visualization in geographical Information Systems*, pages 85–102, Chichester, 1994. John Wiley & Sons.

[26] M. Alexa and W. Müller. Visualization by metamorphosis. In C. M. Wittenbrink and A. Varshney, editors, *IEEE Visualization 1998 Late Breaking Hot Topics Proceedings*, pages 33–36, 1998.

[27] M. Smith, R.J. Taffler, and L White. Cartoon graphics in the communication of accounting information for management decision making. *Journal of Applied Management Accounting Research*, 1(1):31–50, 2002.

[28] R. M. Pickett and G. G. Grinstein. Iconographics displays for visualizing multidimensional data. In *Proc. IEEE Conference on Systems, Man, and Cybernetics*, pages 514–19, 1988.

[29] B. Kleiner and J. Hartigan. Representing points in many dimension by trees and castles. *J. Am. Stat. Ass.*, 76:260–9, 1981.

[30] M. Chua and S. Eick. Information rich glyphs for software management. *IEEE Computer Graphics and Applications*, 18:24–9, 1998.

[31] E. Lieske and R. Myers. *Coral Reef Fishes: Indo-Pacific and Caribbean.* Princeton University Press, 2002.

[32] H. Levkowitz. *Color Theory and Modeling for Computer Graphics, Visualization and Computer Graphics.* Kluwer, Boston, USA, 1997.

[33] S.A. Nene, S.K. Nayar, and H Murase. Columbia object image library (coil-20). Technical report, Columbia University, 1996.